Computer Networks II

Multi Protocol Label Switching

Giorgio Ventre
COMICS LAB
Dipartimento di Informatica e Sistemistica
Università di Napoli Federico II

Nota di Copyright

Quest’insieme di trasparenze è stato ideato e realizzato dai ricercatori del Gruppo di Ricerca sull’Informatica Distribuita del Dipartimento di Informatica e Sistemistica dell’Università di Napoli e del Laboratorio Nazionale per la Informatica e la Telematica Multimediali. Esse possono essere impiegate liberamente per fini didattici esclusivamente senza fini di lucro, a meno di un esplicito consenso scritto degli Autori. Nell’uso dovrà essere esplicitamente riportata la fonte e gli Autori. Gli Autori non sono responsabili per eventuali imprecisioni contenute in tali trasparenze né per eventuali problemi, danni o malfunzionamenti derivanti dal loro uso o applicazione.
**IP: “Best-Effort Philosophy”**

- Well architected, not necessarily worked out in detail
- Realization: can’t predict the future
- Architectural decisions:
  - Make it reasonable
  - Make it flexible
  - Make it extensible

**IP Control Plane Evolution**

- Again, just good enough (best-effort) …
  - But again, flexible, extensible
- Distance Vector routing was fine for quite a while
  - Just in time, along came link state (OSPF and IS-IS)
- Now a burning question in OSPF/IS-IS is:
  - Convergence “in a few seconds” is not good enough?
- **Goal: “Business” IP for service providers…**
  - Make me money – new services, GoS
  - Don’t lose me money – uptime, SLAs
  - OSPF/BGP not originally designed to support QoS or multiple services (eg: VoIP, VPNs)
**ATM - Perfectionist’s Dream**

- Connection-oriented
- Does everything and does it well
- Anticipated all future uses and factored them in
- Philosophical mismatch with IP

**Overlay Model for IP-over-ATM Internetworking**

- Goal: Run IP over ATM core networks
- Why? ATM switches offered performance, predictable behavior and services (CBR, VBR, GFR…)
- ISPs created “overlay” networks that presented a virtual topology to the edge routers in their network
- Using ATM virtual circuits, the virtual network could be reengineered without changing the physical network
- Benefits
  - Full traffic control
  - Per-circuit statistics
  - More balanced flow of traffic across links
Overlay Model (Contd)

- ATM core ringed by routers
- PVCs overlaid onto physical network

![Diagram showing Physical View and Logical View of ATM core with routers and PVCs]

Issue 1: Mapping IP data-plane to ATM: Address Resolution Woes!

- A variety of server-based address resolution servers:
  - ATMARP (RFC 1577), LANE server, BUS server, MPOA server, NHRP server....
  - Use of separate pt-pt and pt-mpt VCs with servers
  - Multiple servers + backup VCs to them needed for fault tolerance
  - Separate servers needed in every LOGICAL domain (eg: LIS)
- Mismatch between the notion of IP subnet and ATM network sizing
  - *Cut-through forwarding* between nodes on same ATM network *hard to achieve!*

![Diagram showing ATMARP servers and routers in LIS domains]

Corso di Reti di Calcolatori II – Anno accademico 2008/2009
Dipartimento di Informatica e Sistemistica, Università di Napoli Federico II
Issue 2: Mapping IP control-plane (e.g. OSPF) to ATM

- Basic OSPF assumes that subnets are pt-pt or offer broadcast capability.
- ATM is a Non-Broadcast Multiple Access (NBMA) media
  - NBMA “segments” support multiple “routers” with pt-pt VCs but do not support data-link broadcast/mcast capability
  - Each VC is costly => setting up full mesh for OSPF Hello messages is prohibitively expensive!

- Two “flooding adjacency” models in OSPF:
  - Non-Broadcast Multiple Access (NBMA) model
  - Point-to-Multipoint (pt-mpt) Model

- Different tradeoffs…

Re-examining Basics: Routing vs Switching

- Routing: Based on address lookup. Max prefix match.
  - Search Operation
  - Complexity ≈ O(log₂n)

- Switching: Based on circuit numbers
  - Indexing operation
  - Complexity O(1)
  - Fast and Scalable for large networks and large address spaces

- These distinctions apply on all datalinks: ATM, Ethernet, SONET
IP Routing vs IP Switching

On ATM networks:
- IP routers use IP addresses
  ⇒ Reassemble IP datagrams from cells
- IP Switches use ATM Virtual circuit numbers
  ⇒ Switch cells
  ⇒ Do not need to reassemble IP datagrams

MPLS: Best of Both Worlds

Note: the "hybrid" also happens to be a solution that bypasses IP-over-ATM mapping woes!
History: Ipsilon’s IP Switching: Concept

- Developed by Ipsilon
- Routing software in every ATM switch in the network
- Initially, packets are reassembled by the routing software and forwarded to the next hop
- Long term flows are transferred to separate VCs. Mapping of VCIs in the switch ⇒ No reassembly

Hybrid: IP routing (control plane) + ATM switching (data plane)

Ipsilon’s IP Switching

ATM VCs setup when new IP “flows” seen, i.e., “data-driven” VC setup

- Flow-oriented traffic: FTP, Telnet, HTTP, Multimedia
- Short-lived Traffic: DNS query, SMTP, NTP, SNMP, request-response Ipsilon claimed that 80% of packets and 90% of bytes are flow-oriented.
- Ipsilon claimed their Generic Switch Management Protocol (GSMP) to be 2000 lines, and Ipsilon Flow Management Protocol (IFMP) to be only 10,000 lines of code
- Runs as added software on an ATM switch
**Ipsilon’s IP Switching Architecture**

**Ipsilon’s IP Switching Protocols**

- **GSMP**
  - General Switch Management Protocol
  - Allows the control module to have access to the switch hardware

- **IFMP**
  - Ipsilon Flow Management Protocol
  - Manages flows and realises associations between IP flows and ATM virtual channels
Issues with Ipsilon’s IP switching

- VCI field is used as ID.
  VPI/VCI change at switch
  ⇒ Must run on every ATM switch
  ⇒ non-IP switches not allowed between IP switches
  ⇒ Subnets limited to one switch
- Cannot support VLANs
- Scalability: Number of VC ≥ Number of flows.
  ⇒ VC Explosion. 1000 setups/sec.
- Quality of service determined implicitly by the flow class or by RSVP
- ATM Only

Tag Switching

- Proposed by CISCO
- Similar to VLAN tags
- Tags can be explicit or implicit L2 header

| L2 Header | Tag |

- Ingress router/host puts a tag. Exit router strips it off.

Key difference: tags can be setup in the background using IP routing protocols (i.e. control-driven VC setup)
Alphabet Soup!

- CSR Cell Switched Router
- ISR Integrated Switch and Router
- LSR Label Switching Router
- TSR Tag Switching Router
- Multi layer switches, Swoters
- DirectIP
- FastIP
- PowerIP

MPLS working group in IETF was formed to reach a common standard

MPLS Broad Concept: Route at Edge, Switch in Core
**MPLS Terminology**

- LDP: Label Distribution Protocol
- LSP: Label Switched Path
- FEC: Forwarding Equivalence Class
- LSR: Label Switching Router
- LER: Label Edge Router (Useful term not in standards)
- MPLS is “multi-protocol” both in terms of the protocols it supports ABOVE it and BELOW it in the protocol stack!

**MPLS Header**

- IP packet is **encapsulated** in MPLS header and sent down LSP
- IP packet is **restored** at end of LSP by egress router
  - TTL is adjusted by default
**MPLS Label Stack Concept**

- Labels = Explicit or implicit L2 header
- TTL = Time to live
- Exp = Experimental
- SI = Stack indicator

Diagram:

- L2 Header
- Label Stack Entry
- Label Stack Entry

20b 3b 1b 8b

Label Exp SI TTL

Allows nested tunnels, that are opaque, i.e. do not know or care what protocol data they carry (a.k.a multi-protocol)

---

**MPLS Header**

- **Label**
  - Used to match packet to LSP
- **Experimental bits**
  - Carries packet queuing priority (CoS)
- **Stacking bit**: can build “stacks” of labels
  - Goal: nested tunnels!
- **Time to live**
  - Copied from IP TTL
Multi-protocol operation

ATM
Frame Relay
TDM
X25

Frequency or Wavelength substitution

G-MPLS

Label Encapsulation

L2 Label
VPI VCI DLCI “Shim Label”

“Shim Label” ……

IP | PAYLOAD

MPLS Encapsulation is specified over various media types. Top labels may use existing format, lower label(s) use a new “shim” label format.
MPLS Encapsulation - ATM

ATM LSR constrained by the cell format imposed by existing ATM standards

ATM Header Format

- VPI
- VCI
- PT
- CLP
- HEC

Option 1: Label
Option 2: Combined Label
Option 3: ATM VPI (Tunnel)

5 Octets

ATM Header

- Generic Label Encap. (PPP/LAN format)
- Network Layer Header and Packet (e.g., IP)

AAL 5 PDU Frame (nx48 bytes)

AAL5 Trailer

ATM SAR

1. Top 1 or 2 labels are contained in the VPI/VCI fields of ATM header
2. Further fields in stack are encoded with 'shim' header in PPP/LAN format
   - must be at least one, with bottom label distinguished with 'explicit NULL'
3. TTL is carried in top label in stack, as a proxy for ATM header (that lacks TTL)

MPLS Encapsulation - Frame Relay

Q.922 Header

- Generic Encap. (PPP/LAN Format)

Layer 3 Header and Packet

DLCI: 10, 17, 23 Bits

- Current label value carried in DLCI field of Frame Relay header
- Can use either 2 or 4 octet Q.922 Address (10, 17, 23 bytes)
- Generic encapsulation contains n labels for stack of depth n
  - top label contains TTL (which FR header lacks), 'explicit NULL' label value
**MPLS Encapsulation: PPP & LAN Data Links**

- **Label Stack Entry Format**
  - Label: Label Value, 20 bits (0-16 reserved)
  - Exp.: Experimental, 3 bits (was Class of Service)
  - S: Bottom of Stack, 1 bit (1 = last entry in label stack)
  - TTL: Time to Live, 8 bits

- **Layer 2 Header** (eg. PPP, 802.3)
- **Network Layer Header and Packet** (eg. IP)

- **4 Octets**

- **MPLS ‘Shim’ Headers** (1-n)

- **MPLS on PPP links and LANs** uses ‘Shim’ Header Inserted Between Layer 2 and Layer 3 Headers

- **MPLS Forwarding: Example**
  - An IP packet destined to 134.112.1.5/32 arrives in SF
  - San Francisco has route for 134.112/16
    - Next hop is the LSP to New York
MPLS Forwarding Example

- San Francisco pre-pends MPLS header onto IP packet and sends packet to first transit router in the path.

Because the packet arrived at Santa Fe with an MPLS header, Santa Fe forwards it using the MPLS forwarding table.
- MPLS forwarding table derived from mpls.0 switching table.
**MPLS Forwarding Example**

- Packet arrives from penultimate router with label 0
- Egress router sees label 0 and strips MPLS header
- Egress router performs standard IP forwarding decision

**Label Setup/Signaling: MPLS Using IP Routing Protocols**

- Destination based forwarding tables as built by OSPF, IS-IS, RIP, etc.
Regular IP Forwarding

IP destination address unchanged in packet header!

MPLS Label Distribution

Mapping: 0.40

Request: 47.1
- A Vanilla LSP is actually part of a tree from every source to that destination (unidirectional).
- Vanilla LDP builds that tree using existing IP forwarding tables to route the control messages.
ER-LSP follows route that source chooses. In other words, the control message to establish the LSP (label request) is source routed.
ER LSP - advantages

- Operator has routing flexibility (policy-based, QoS-based)
- Can use routes other than shortest path
- Can compute routes based on constraints in exactly the same manner as ATM based on distributed topology database.
  (traffic engineering)

ER LSP - open problems

- Two signaling options proposed in the standards:
  CR-LDP, RSVP extensions:
  — CR-LDP = LDP + Explicit Route
  — RSVP ext = Traditional RSVP + Explicit Route + Scalability Extensions
- Today, used both options